GUI for the msmsTests package

Label-free SpC LC-MS/MS differential expression

Josep Gregori, Alex Sanchez, and Josep Villanueva
Vall Hebron Institute of Oncology &
Statistics Dept. Barcelona University
josep.gregori@gmail.com

January 5, 2014

1 Introduction

The graphical user interface (GUI) described in this document uses the functions in the freely available Bioconductor R package `msmsTests` [1] for inference in label-free protein differential expression with spectral counts (SpC) [2] [3].

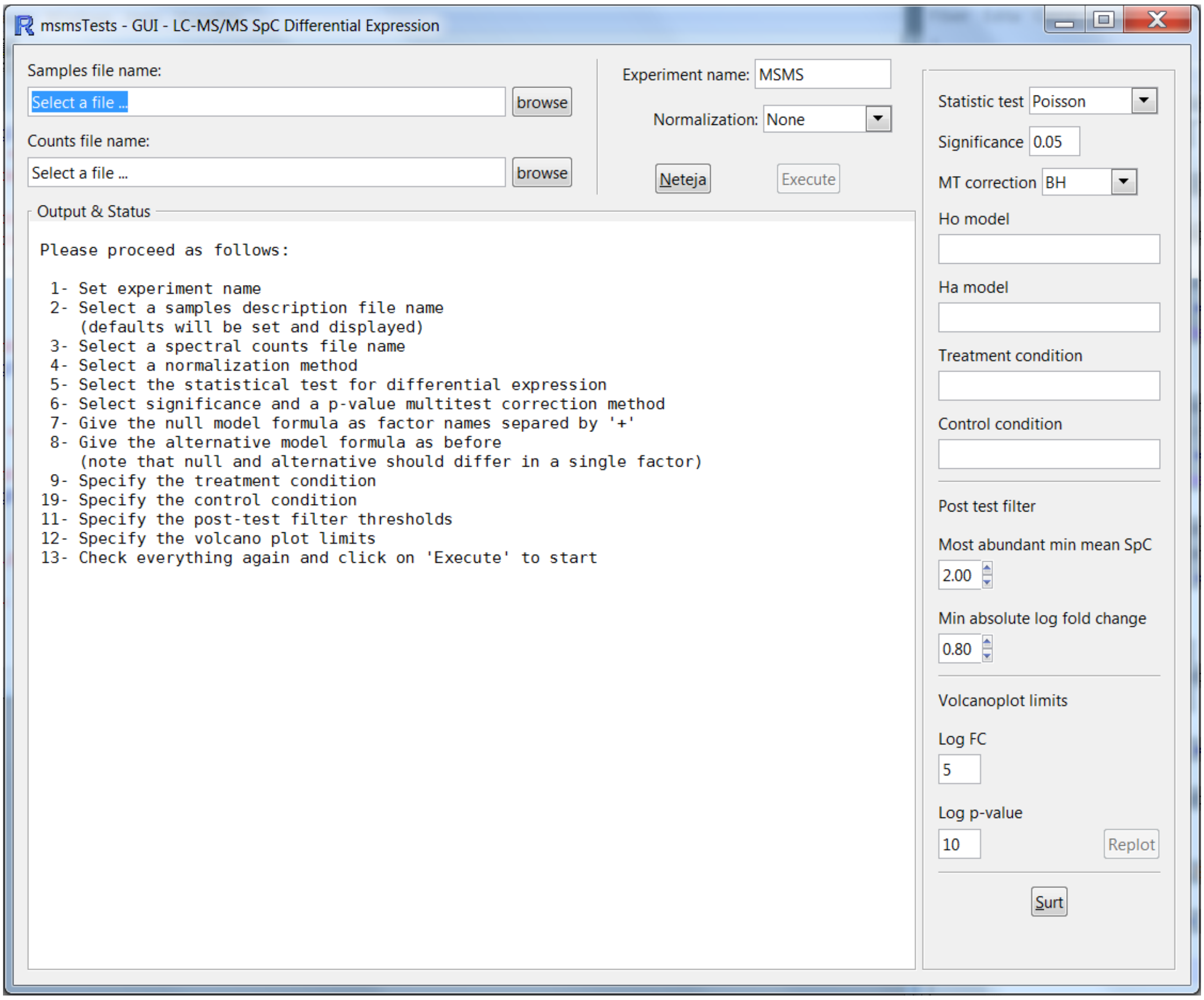
The tests are GLM regression based [4], with methods for the Poisson or the negative binomial distributions, or the quasi-likelihood GLM extension. The linear model may include a treatment factor and eventual blocking factors [6]. To improve the reproducibility [5] of the list of differentially expressed proteins (DEP), a post-test filter may be applied to the results of the tests to flag those proteins with a minimum signal and effect size.

This document does not describe the methods but the uses of the GUI, and some background on label-free SpC differential proteomics is supposed. For further help see the `msmsTests` manual and vignettes at

http://www.bioconductor.org/packages/release/bioc/html/msmsTests.html.

A companion document "*msmsEDA and msmsTests: R/Bioconductor packages for spectral count label-free proteomics data analysis*" by the same authors, is an extra source of help. The document describes the context and the challenges in proteomics biomarker discovery, and the use of the functions in these packages to solve them. A set of results illustrate the advantages in the use of blocking factors when batch effects are evidenced, and in the use of the post-test filter to improve the reproducibility of DEP lists.

2 The options



msmsTests - GUI - LC-MS/MS SpC Differential Expression

Samples file name: Select a file ...

Counts file name: Select a file ...

Experiment name: MSMS

Normalization: None

Output & Status

```
Please proceed as follows:

 1- Set experiment name
 2- Select a samples description file name
    (defaults will be set and displayed)
 3- Select a spectral counts file name
 4- Select a normalization method
 5- Select the statistical test for differential expression
 6- Select significance and a p-value multitest correction method
 7- Give the null model formula as factor names separed by '+'
 8- Give the alternative model formula as before
    (note that null and alternative should differ in a single factor)
 9- Specify the treatment condition
19- Specify the control condition
11- Specify the post-test filter thresholds
12- Specify the volcano plot limits
13- Check everything again and click on 'Execute' to start
```

Statistic test: Poisson

Significance: 0.05

MT correction: BH

Ho model:

Ha model:

Treatment condition:

Control condition:

Post test filter

Most abundant min mean SpC: 2.00

Min absolute log fold change: 0.80

Volcanoplot limits

Log FC: 5

Log p-value: 10

Figure 1: The graphical user interface.

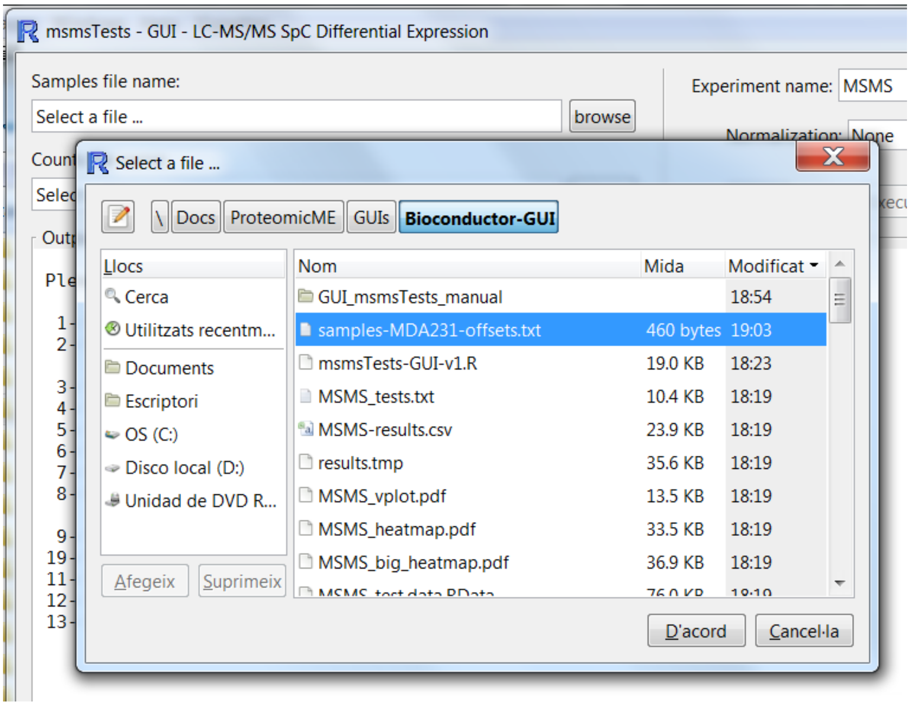
Figure 1 shows the appearance of the `msmsTests` GUI when just started its execution. The output window shows to the user a note of help, with the recommended order of actions. According to this recommended order we shall:

1. **Give an experiment name**

   This name will be used as root for the different file names with figures, tables and reports produced with the results.

2. **Select a file with the samples description**

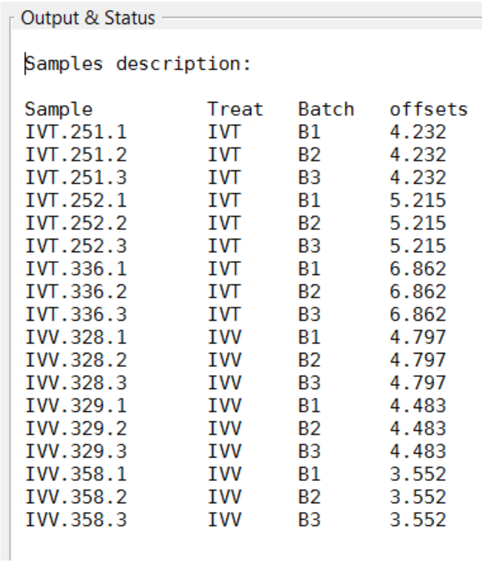
   This file must be a tab delimited text file with different columns and a header. The header of the first column must be 'Samples' and give the IDs of each sample. Subsequent columns will be interpreted as factors with the corresponding levels of each sample. The header of each factor column is the factor name to be used in the null and alternative models. An extra column with header 'offsets' may be used and will be interpreted as normalization factors.



Nom	Mida	Modificat
GUI_msmsTests_manual		18:54
samples-MDA231-offsets.txt	460 bytes	19:03
msmsTests-GUI-v1.R	19.0 KB	18:23
MSMS_tests.txt	10.4 KB	18:19
MSMS-results.csv	23.9 KB	18:19
results.tmp	35.6 KB	18:19
MSMS_vplot.pdf	13.5 KB	18:19
MSMS_heatmap.pdf	33.5 KB	18:19
MSMS_big_heatmap.pdf	36.9 KB	18:19
MSMS_test_data.RData	76.0 KB	18:19

Figure 2: Select a samples description file.

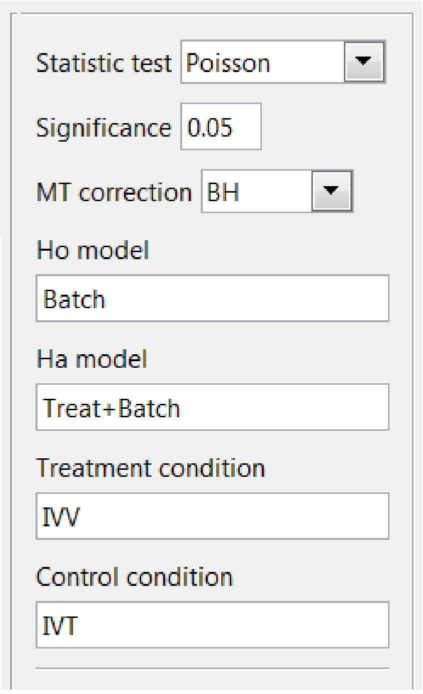
Once selected, the file will be read and its contents displayed for the user to check it. In response to the treatment of the samples description file a number of defaults will be set and displayed in the corresponding GUI controls, as shown in Figure 4.



```
Samples description:

Sample        Treat   Batch   offsets
IVT.251.1     IVT     B1      4.232
IVT.251.2     IVT     B2      4.232
IVT.251.3     IVT     B3      4.232
IVT.252.1     IVT     B1      5.215
IVT.252.2     IVT     B2      5.215
IVT.252.3     IVT     B3      5.215
IVT.336.1     IVT     B1      6.862
IVT.336.2     IVT     B2      6.862
IVT.336.3     IVT     B3      6.862
IVV.328.1     IVV     B1      4.797
IVV.328.2     IVV     B2      4.797
IVV.328.3     IVV     B3      4.797
IVV.329.1     IVV     B1      4.483
IVV.329.2     IVV     B2      4.483
IVV.329.3     IVV     B3      4.483
IVV.358.1     IVV     B1      3.552
IVV.358.2     IVV     B2      3.552
IVV.358.3     IVV     B3      3.552
```

Figure 3: Samples description file just read.

The defaults are the null and the alternative model, and the two levels of the factor to contrast. The default null model is taken as the set of factors excluding the first. The alternative model is taken as the set of all factors given. So by default the first factor in the samples description file is taken as the ’treatment’ factor, the other factors are interpreted as blocking factors. The treatment and control levels of the treatment factor are taken by default as the first and second. These defaults may be changed as needed.



Statistic test Poisson

Significance 0.05

MT correction BH

Ho model

Batch

Ha model

Treat+Batch

Treatment condition

IVV

Control condition

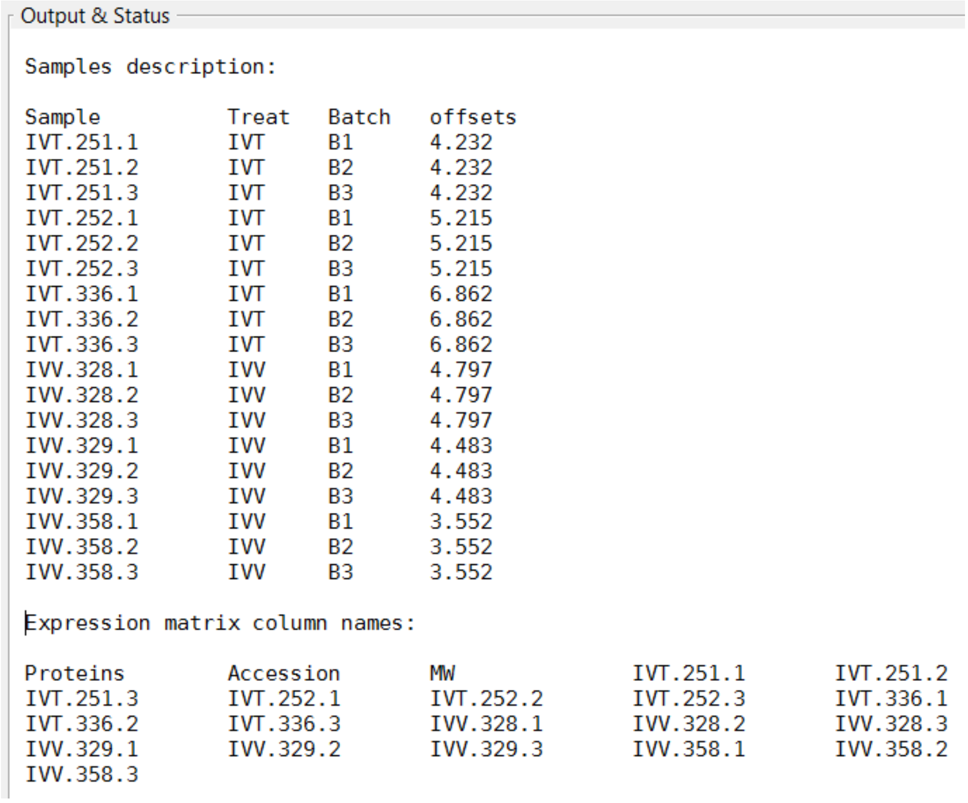
IVT

Figure 4: Defaults set in response to the samples description given.

3. **Select a file with the SpC table, and proteins description.**

   This file must be a tab delimited text file with a header. A column with header ’Accession’ with the proteins IDs is required, and will be used to name the proteins in the results. Each sample in the samples description table must have a column in this file, whose header must be the same ID. Extra columns in the SpC table are harmless and will be ignored. An optional column with header ’Proteins’ will be interpreted as a textual description of each protein, including a field of the form ”GN=[A-Z0-9_]*” from where the gene name will be taken.

   The column names in the SpC table file will be displayed in the GUI output control for the user to check.

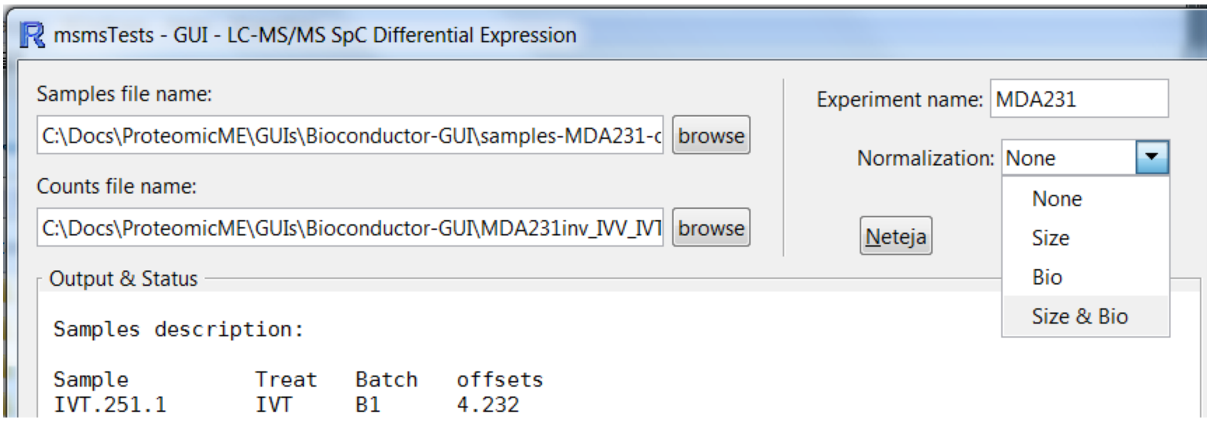


```
Samples description:

Sample           Treat    Batch    offsets
IVT.251.1        IVT      B1       4.232
IVT.251.2        IVT      B2       4.232
IVT.251.3        IVT      B3       4.232
IVT.252.1        IVT      B1       5.215
IVT.252.2        IVT      B2       5.215
IVT.252.3        IVT      B3       5.215
IVT.336.1        IVT      B1       6.862
IVT.336.2        IVT      B2       6.862
IVT.336.3        IVT      B3       6.862
IVV.328.1        IVV      B1       4.797
IVV.328.2        IVV      B2       4.797
IVV.328.3        IVV      B3       4.797
IVV.329.1        IVV      B1       4.483
IVV.329.2        IVV      B2       4.483
IVV.329.3        IVV      B3       4.483
IVV.358.1        IVV      B1       3.552
IVV.358.2        IVV      B2       3.552
IVV.358.3        IVV      B3       3.552

Expression matrix column names:

Proteins         Accession        MW               IVT.251.1        IVT.251.2
IVT.251.3        IVT.252.1        IVT.252.2        IVT.252.3        IVT.336.1
IVT.336.2        IVT.336.3        IVV.328.1        IVV.328.2        IVV.328.3
IVV.329.1        IVV.329.2        IVV.329.3        IVV.358.1        IVV.358.2
IVV.358.3
```

Figure 5: List of variables in the SpC file just read.

4. **Select a normalization method**

   The normalization is implemented by means of offsets in the GLM [4]. The methods offered are: i) *None*, for no normalization. ii) *Size*, normalizing by the total SpC by sample. iii) *Bio*, the offsets are taken from the 'offsets' column in the samples description table. And iv) *Size & Bio*, the normalization factors are computed as the product of the total SpC by sample and the offsets in the samples description table. Normalization by *Size* assumes that the total protein measured in each sample is the same, so that the expected total SpC should be alse the same. Normalization by *Bio* considers a normalization with some biological meaning, as in a cell-to-cell comparison. Normalization by *Size & Bio* assumes both that the total protein measured in each sample is the same and a normalization with some biological meaning.



msmsTests - GUI - LC-MS/MS SpC Differential Expression

Samples file name: C:\Docs\ProteomicME\GUIs\Bioconductor-GUI\samples-MDA231-c

Counts file name: C:\Docs\ProteomicME\GUIs\Bioconductor-GUI\MDA231inv_IVV_IVT

Experiment name: MDA231

Normalization: None

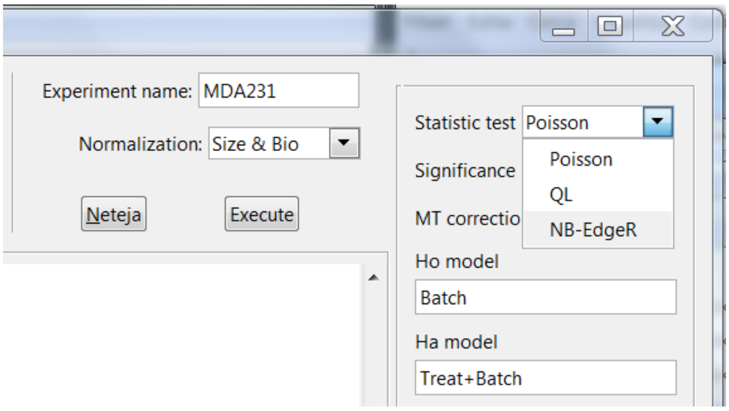
Output & Status

```
Samples description:

Sample          Treat   Batch   offsets
IVT.251.1       IVT     B1      4.232
```

Figure 6: Select a normalization method.

5. **Select a test**

   Each row of the SpC matrix, corresponding to a single protein, is tested by the likelihood ratio comparing the null and the alternative models fitted by GLM regression. The methods available are: i) *Poisson* using the Poisson distribution. ii) *QL* using the quasi-likelihood extension to the GLM. And iii) *NB-EdgeR* using the negative binomial implementation in the package `edgeR` [7]. The Poisson GLM is indicated when only technical replicates are available. With just two or three biological replicates for each condition, the NB GLM of `edgeR` is the best option. Above four replicates, both the NB or the QL could give good results.



Experiment name: MDA231

Normalization: Size & Bio

Statistic test: Poisson

Ho model: Batch

Ha model: Treat+Batch

Figure 7: Select a test method.

6. **Select significance and a multitest p-value correction method**

   By default the significance is set to 0.05, and may be modified. The only multitest method currently available if Benjamini-Hochberg [8]

7. **Specify a null model**

Beyond the default any null model may be specified. A model with just an intercept is specified as '1'. Alternatively a null model may be specified as a set of blocking factors separated by '+'.

8. **Specify an alternative model**

   Beyond the default any alternative model may be specified. Null and alternative must differ in a single treatment factor, given in the alternative but not in the null (See Figure 7).

9. **Specify the treatment level**

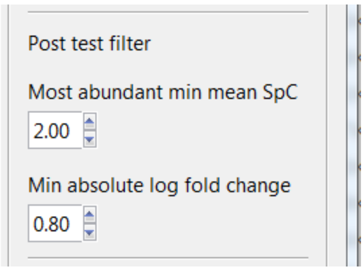
   This must be a level of the treatment factor.

10. **Specify the control level**

    This must be a level of the treatment factor. The FC will be computed as the ratio of the estimated expression treatment/control.

11. **Specify the post-test filter thresholds**

    When reproducibility is of concern the DEPs should not rely solely in p-values. A minimum of signal and effect size is required. The level of this minimum depends of multiple factors. We recommend setting a minimum of 2 SpC in the most abundant condition, and a minimum absolute log fold change of 0.8. When the list of DEPs is very long, higher thresholds are recommended. Instead for a short list of DEPs, lower values could be used, although this is not recommended.



Post test filter

Most abundant min mean SpC

2.00

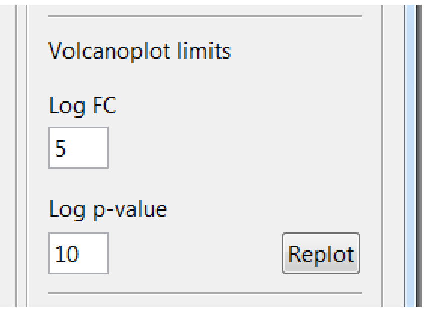
Min absolute log fold change

0.80

Figure 8: Post-test filter options to improve reproducibility.

12. **Specify the volcano plot axis limits**

    Often very few DEPs with very low p-values produce distorted volcano plots, where most of the information is compresses down. Actually the most relevant information in a volcano plot is around the limits of the p-value and LogFC thresholds. To help in the obtaining of a meaningful plot the limits of both axis may be specified. Default values are ±5 for the LogFC in the x axis, and 10 for the $-log_{10} pvalue$ in the y axis.



Volcanoplot limits

Log FC: 5

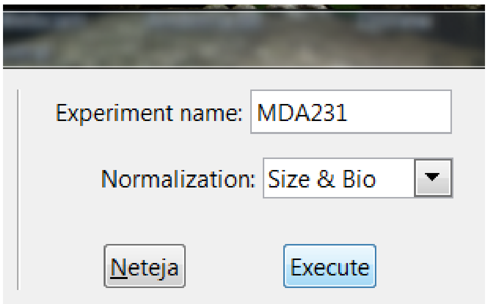
Log p-value: 10

Figure 9: Volcanoplot axis limits.

The volcano plot is sent to a pdf file, where it may be checked. When the test execution has finished the button 'Replot' is enabled, permitting to redefine new axis limits and to redraw the volcano plot.

13. **Check everything and click on the 'Exec' button**

When both the samples description table and the SpC table have been loaded, after item 3 in our list of recommended actions, the 'Exec' button becomes enabled because all relevant information is already available with the defaults. Items 4 to 12 are used to check or edit these defaults. At this point everything is ready and after rechecking all info we may click on the 'Exec' button to start the computations according to our settings.

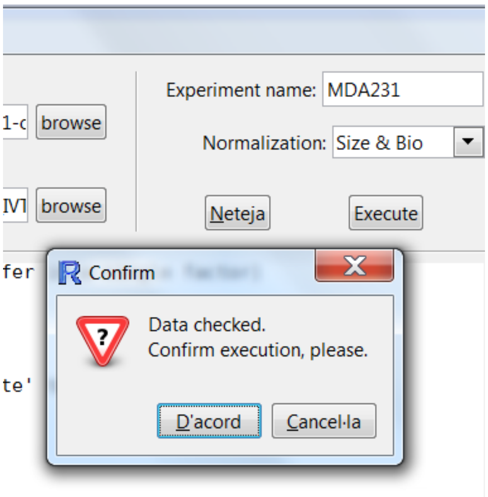


Experiment name: MDA231

Normalization: Size & Bio

Figure 10: Click on the execute button to start the computation.

The first step in the computation is checking that all parameters conform. If they don't an error message is given and the computation is stopped. Otherwise a dialog is displayed asking for the confirmation to proceed.



Experiment name: MDA231

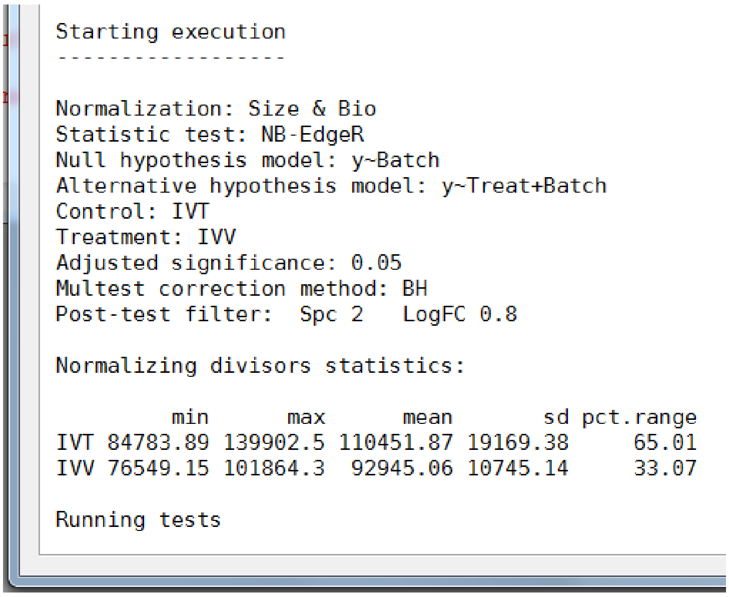
Normalization: Size & Bio

Data checked.
Confirm execution, please.

Figure 11: With parameters checked and OK, we are asked to confirm execution in a dialog control.

3 The results

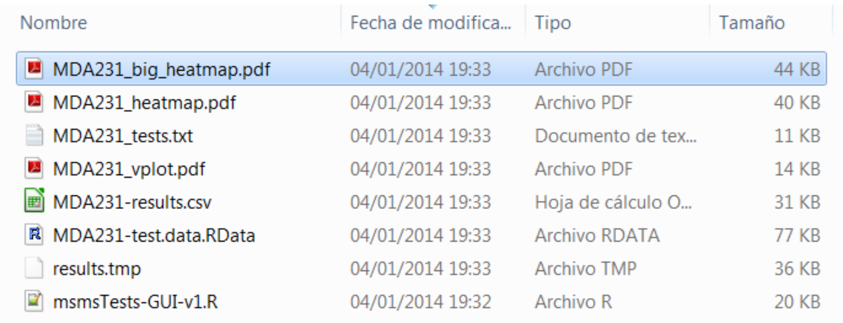
When the computations start the first we see is a summary with all the selected parameters, and statistics of the given normalizing factors, as in Figure 12. The message 'Running tests' warns us that some time is required to obtain the results. In case of any runtime error, a message is displayed in the main R window. Otherwise after a while we get a summary of the results in the GUI output control.



```
Starting execution
------------------

Normalization: Size & Bio
Statistic test: NB-EdgeR
Null hypothesis model: y~Batch
Alternative hypothesis model: y~Treat+Batch
Control: IVT
Treatment: IVV
Adjusted significance: 0.05
Multest correction method: BH
Post-test filter:  Spc 2   LogFC 0.8

Normalizing divisors statistics:

         min      max      mean       sd pct.range
IVT 84783.89 139902.5 110451.87 19169.38     65.01
IVV 76549.15 101864.3  92945.06 10745.14     33.07

Running tests
```

Figure 12: Methods and conditions being used in the computations.

A number of files are produced with the results. These files have names starting by the experiment name and followed by descriptive names and extensions. They are as in Figure 13:



Nombre	Fecha de modifica...	Tipo	Tamaño
MDA231_big_heatmap.pdf	04/01/2014 19:33	Archivo PDF	44 KB
MDA231_heatmap.pdf	04/01/2014 19:33	Archivo PDF	40 KB
MDA231_tests.txt	04/01/2014 19:33	Documento de tex...	11 KB
MDA231_vplot.pdf	04/01/2014 19:33	Archivo PDF	14 KB
MDA231-results.csv	04/01/2014 19:33	Hoja de cálculo O...	31 KB
MDA231-test.data.RData	04/01/2014 19:33	Archivo RDATA	77 KB
results.tmp	04/01/2014 19:33	Archivo TMP	36 KB
msmsTests-GUI-v1.R	04/01/2014 19:32	Archivo R	20 KB

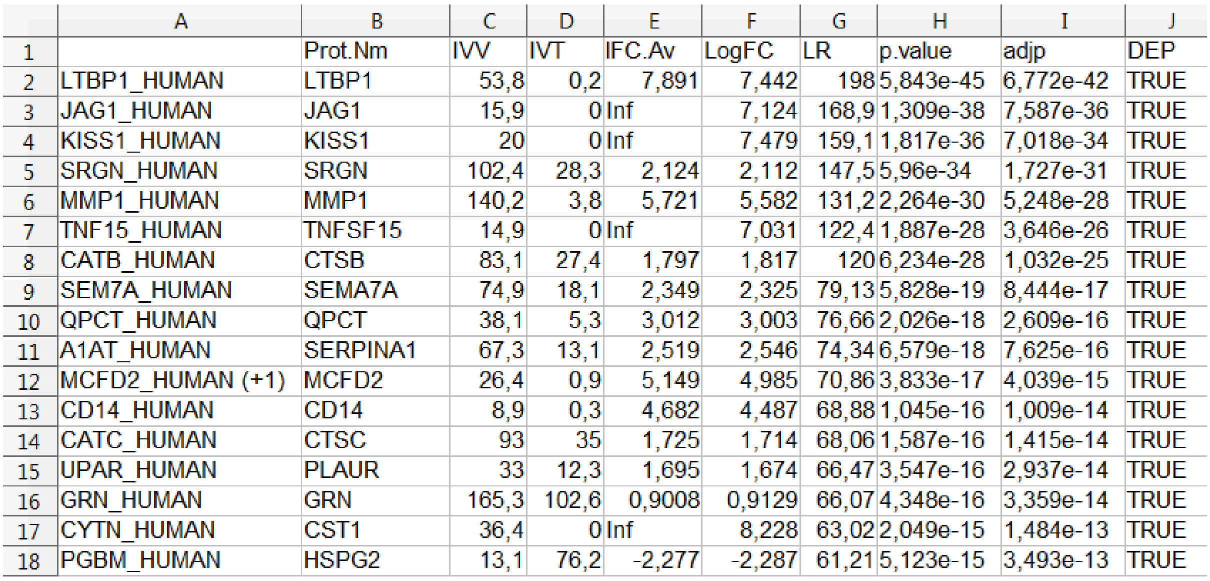
Figure 13: Files generated with the results

- **A text file with the output**

  A text file ending with ’_tests.txt’ with all the output sent to the GUI output control. That is i) the samples description table. ii) the column names in the original SpC table. iii) the list of parameters values. iv) a summary with statistics of the normalizing factors. v) a cross-tabulation of the number of features by bins of p-values and LogFC with. vi) the number of features with p-value below significance. vii) the number of significant features passing the post-test filter. And viii) a table with the results of the top 100 features in ascending order of p-value.

- **A tab delimited file with the table of results**

  The table of DEPs, the proteins with multitest adjusted p-vales below the significance, is given in a tab delimited file with name ending as ’-results.csv’. This file may be imported into an spread sheet program or directly displayed as a text file (See Figure 14). The first column contains the Accessions. If the SpC table contained a ’Proteins’ column with descriptions and gene names, the gene names are given in the second column. Then follow the observed average SpC for each of the two conditions compared. The ’lFC.Av’ as the LogFC computed from the SpC averages taking into account the normalizing factors. The ’LogFC’ as the LogFC estimated from the fitted model. The ’LR’ as the statistic obtained in the comparison. The ’p.value’ as the raw p-value in the test. The ’adjp’ as the BH multitest adjusted p-value. And the ’DEP’ with flags TRUE or FALSE according to the used reproducibility criteria. When there are no blocking factors the values ’lFC.Av’ and ’LogFC’ will be the same. When there are blocking factors they may differ.

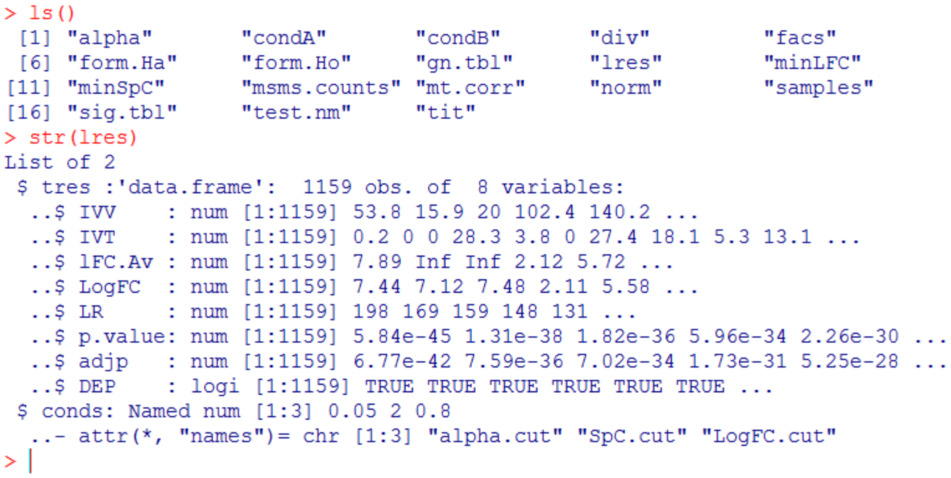


	A	B	C	D	E	F	G	H	I	J
1		Prot.Nm	IVV	IVT	lFC.Av	LogFC	LR	p.value	adjp	DEP
2	LTBP1_HUMAN	LTBP1	53,8	0,2	7,891	7,442	198	5,843e-45	6,772e-42	TRUE
3	JAG1_HUMAN	JAG1	15,9	0	Inf	7,124	168,9	1,309e-38	7,587e-36	TRUE
4	KISS1_HUMAN	KISS1	20	0	Inf	7,479	159,1	1,817e-36	7,018e-34	TRUE
5	SRGN_HUMAN	SRGN	102,4	28,3	2,124	2,112	147,5	5,96e-34	1,727e-31	TRUE
6	MMP1_HUMAN	MMP1	140,2	3,8	5,721	5,582	131,2	2,264e-30	5,248e-28	TRUE
7	TNF15_HUMAN	TNFSF15	14,9	0	Inf	7,031	122,4	1,887e-28	3,646e-26	TRUE
8	CATB_HUMAN	CTSB	83,1	27,4	1,797	1,817	120	6,234e-28	1,032e-25	TRUE
9	SEM7A_HUMAN	SEMA7A	74,9	18,1	2,349	2,325	79,13	5,828e-19	8,444e-17	TRUE
10	QPCT_HUMAN	QPCT	38,1	5,3	3,012	3,003	76,66	2,026e-18	2,609e-16	TRUE
11	A1AT_HUMAN	SERPINA1	67,3	13,1	2,519	2,546	74,34	6,579e-18	7,625e-16	TRUE
12	MCFD2_HUMAN (+1)	MCFD2	26,4	0,9	5,149	4,985	70,86	3,833e-17	4,039e-15	TRUE
13	CD14_HUMAN	CD14	8,9	0,3	4,682	4,487	68,88	1,045e-16	1,009e-14	TRUE
14	CATC_HUMAN	CTSC	93	35	1,725	1,714	68,06	1,587e-16	1,415e-14	TRUE
15	UPAR_HUMAN	PLAUR	33	12,3	1,695	1,674	66,47	3,547e-16	2,937e-14	TRUE
16	GRN_HUMAN	GRN	165,3	102,6	0,9008	0,9129	66,07	4,348e-16	3,359e-14	TRUE
17	CYTN_HUMAN	CST1	36,4	0	Inf	8,228	63,02	2,049e-15	1,484e-13	TRUE
18	PGBM_HUMAN	HSPG2	13,1	76,2	-2,277	-2,287	61,21	5,123e-15	3,493e-13	TRUE

Figure 14: The csv file produced with the results

- **A RData file**

  A Rdata file with all relevant R objects (See Figure 15). *lres* is a list with two items, the full table of results and the post-test filter conditions. *sig.tbl* contains a cross table with number of features by bins of p-value and LogFC. *msms.counts* contains the SpC matrix as used in the tests, excluding all non relevant columns. *gn.tbl* is the list of gene names, whose names are the accessions. *div* is the vector of normalizing factors by sample. The remaining objects contain the parameters used in the computations.



```
> ls()
 [1] "alpha"       "condA"       "condB"       "div"         "facs"
 [6] "form.Ha"     "form.Ho"     "gn.tbl"      "lres"        "minLFC"
[11] "minSpC"      "msms.counts" "mt.corr"     "norm"        "samples"
[16] "sig.tbl"     "test.nm"     "tit"
> str(lres)
List of 2
 $ tres :'data.frame':  1159 obs. of  8 variables:
  ..$ IVV    : num [1:1159] 53.8 15.9 20 102.4 140.2 ...
  ..$ IVT    : num [1:1159] 0.2 0 0 28.3 3.8 0 27.4 18.1 5.3 13.1 ...
  ..$ lFC.Av : num [1:1159] 7.89 Inf Inf 2.12 5.72 ...
  ..$ LogFC  : num [1:1159] 7.44 7.12 7.48 2.11 5.58 ...
  ..$ LR     : num [1:1159] 198 169 159 148 131 ...
  ..$ p.value: num [1:1159] 5.84e-45 1.31e-38 1.82e-36 5.96e-34 2.26e-30 ...
  ..$ adjp   : num [1:1159] 6.77e-42 7.59e-36 7.02e-34 1.73e-31 5.25e-28 ...
  ..$ DEP    : logi [1:1159] TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ conds: Named num [1:3] 0.05 2 0.8
  ..- attr(*, "names")= chr [1:3] "alpha.cut" "SpC.cut" "LogFC.cut"
> |
```

Figure 15: The objects in the .RData file produced, with the results

- **Volcano plot**

A volcano plot is sent to the file ending with '_vplot.pdf'. After running the tests the button 'Replot' (See Figure 9) is enabled, so that the axis limits may be changed and the volcano plot replotted accordingly. The plot will be sent to the same pdf file, changing the name of the previous pdf will preserve its contents.

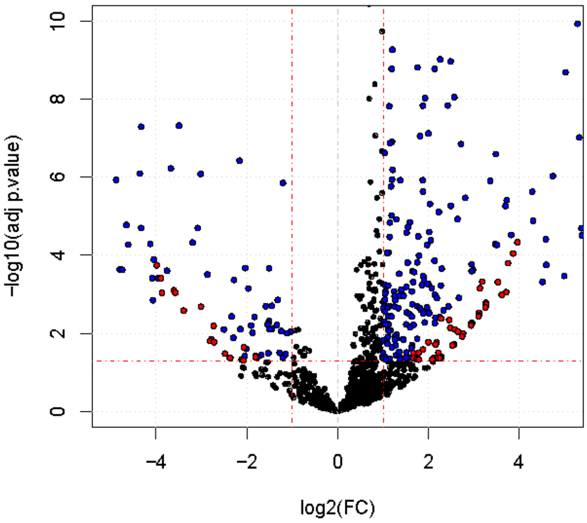


Figure 16: The volcanoplot sent to a pdf file

The black dots are features with adjusted p-value above significance, or with LogFC below the post-test filter threshold. The red dots are proteins with p-value below significance, LogFC above the threshold in the post-test filter, but with SpC in the most abundant condition below the threshold in the post-test filter. Blue dots are the declared DEPs, that is significant proteins with signal and effect size above thresholds in the post-test filter.

- **Heatmap**

  A heatmap with all DEPs is sent to the file ending with '_heatmap.pdf'. The signal used is the raw SpC divided by the corresponding normalization factor, which is further centered and scaled to 1 sd for each DEP. See Figure 17.

- **Expanded heatmap**

  An expanded heatmap with all DEPs is sent to the file ending with '_big_heatmap.pdf'. It has 3mm height by feature so that the accessions may be easily read. As before, the signal used is the raw SpC divided by the corresponding normalization factor, and further centered and scaled by DEP. Although a heatmap on DEPs has no discovery value, it is useful in visualizing the consistency of the expression in each condition. This allows to visualize and identify the expected level of reproducibility of each DEP, as the colors give an idea of distribution overlap. With blocking factors the signal will be masked by its influence. See Figure 18.

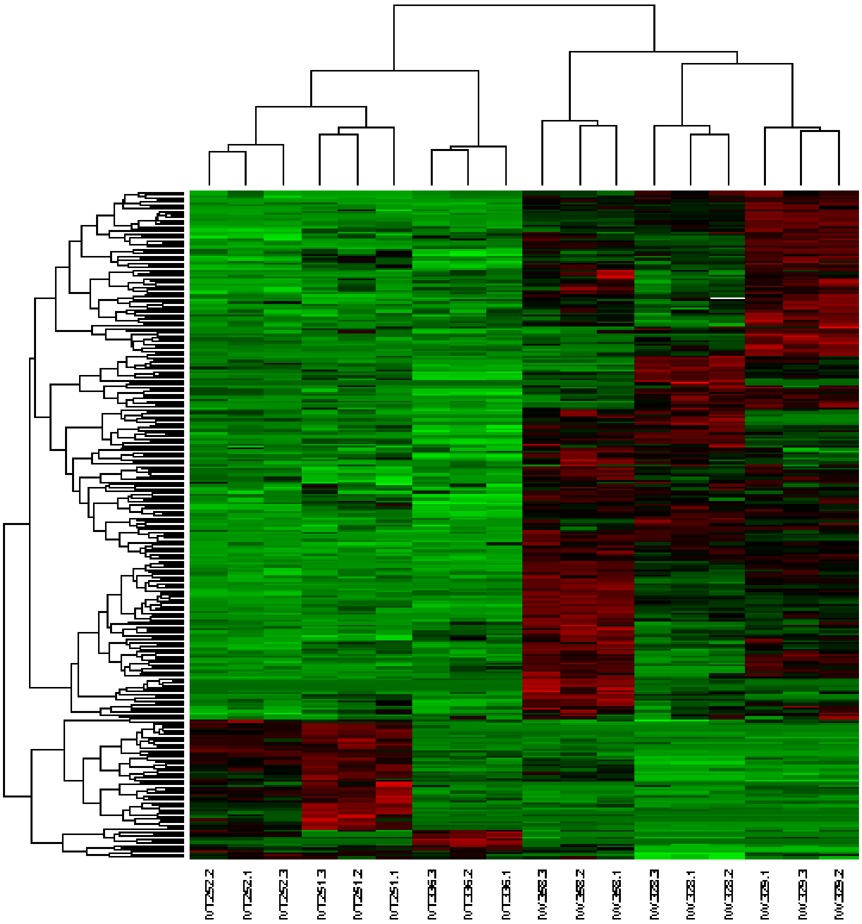


Figure 17: Full heatmap with the DEPs set sent to a pdf file

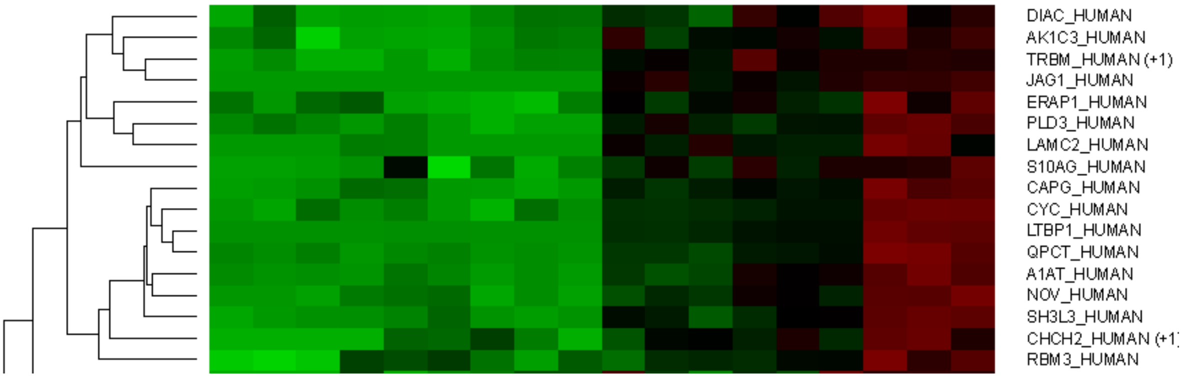


Figure 18: Detail of the expanded heatmap, with accession names

4 Again

The button 'Exit' will terminate the GUI. In case that a new computation is required an option is to click on the 'Clear' button which will reset all variables. Then start anew the list of recommended actions.

References

[1] Gregori, J. et al. *msmsTests: LC-MS/MS Differential Expression Tests.*
R package version 1.0.0.
http://www.bioconductor.org/packages/release/bioc/html/msmsTests.html

[2] Mallick P., Kuster B. *Proteomics: a pragmatic perspective.* Nat Biotechnol 2010;28:695-709.

[3] Neilson K.A., Ali N.A., Muralidharan S., Mirzaei M., Mariani M., Assadourian G., et al. *Less label, more free: approaches in label-free quantitative mass spectrometry.* Proteomics 2011;11:535-53.

[4] Agresti A., *Categorical Data Analysis*, Wiley-Interscience, Hoboken NJ, 2002

[5] Gregori J., Villareal L., Sanchez A., Baselga J., Villanueva J., *An Effect Size Filter Improves the Reproducibility in Spectral Counting-based Comparative Proteomics.* Journal of Proteomics 2013, http://dx.doi.org/10.1016/j.jprot.2013.05.030

[6] Gregori J., Villareal L., Mendez O., Sanchez A., Baselga J., Villanueva J., *Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics*, Journal of Proteomics, 2012, 75, 3938-3951

[7] Robinson MD, McCarthy DJ and Smyth GK (2010). *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics 26, 139-140

[8] Benjamini, Y., and Hochberg, Y. (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society Series B, 57, 289-300.